Available online at www.sciencedirect.com

SCIENCE @ DIRECT®

Computers, Environment and Urban Systems
30 (2006) 460–484

Computers, Environment and Urban Systems

www.elsevier.com/locate/compenvurbsys

Combining mining and visualization tools to discover the geographic structure of a domain

Josiane Mothe *, Claude Chrisment, Taoufiq Dkaki, Bernard Dousset, Saïd Karouach

Institut de Recherche en Informatique de Toulouse, 118 Route de Narbonne, F-31062 Toulouse Cedex 04, France

Received 27 October 2004; received in revised form 12 September 2005; accepted 12 September 2005

Abstract

Science monitoring is a core issue in the new world of business and research. Companies and institutes need to monitor the activities of their competitors, get information on the market, changing technologies or government policies. This paper presents the Tétralogie platform that is aimed at allowing a user to interactively discover trends in scientific research and communities from large textual collections that include information about geographical location. Tétralogie consists of several agents that communicate with each other on users' demands in order to deliver results to them. Metadata and document content are extracted before being mined. Results are displayed in the form of histograms, networks and geographical maps; these complementary types of presentations increase the possibilities of analysis compared to the use of these tools separately. We illustrate the overall process through a case study of scientific literature analysis and show how the different agents can be combined to discover the structure of a domain. The system correctly predicts the country contribution to a field in future years and allows exploration of the relationships between countries.

Keywords: Information mining; Data analysis; Domain knowledge; Knowledge discovery; Information visualization; Geographical maps

* Corresponding author.
E-mail address: mothe@irit.fr (J. Mothe).

0198-9715/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compenvurbsys.2005.09.004

1. Introduction

Competitive intelligence is a core issue in the new world of business and research. Companies and institutes need to monitor the activities of their competitors, identify information on the market, technologies, or government actions. These monitoring activities are necessary for them to define alliance strategies, innovation and customer oriented strategies. Organizations need methods and tools to lead such activities, that gather information, mine them and display the results in a friendly and efficient way. Large-scale analysis becomes possible thanks to the availability of large sources of publication, patent, scientific literature (Buter & Noyons, 2002), and other data available in electronic form.

Analyzing scientific publications to discover trends and to know the structure of a scientific field and the evolution of scientific communities or topics have been widely explored in the literature, specifically, but not exclusively, in scientometrics (Leydesdorff, 1995). Different types of analysis can be done. In information science, citation and co-citation analysis have been studied in the past as a way to monitor scientific activities (White, 2003; White & McCain, 1998). Citation analysis is used to identify core groups of publications, authors and journals. ISI Web of KnowledgeSM®[1] for example uses citation analysis to determine the history of journal citation and authors. In the same way, hypertext references are mined in the web context, to determine the authority of the pages and to re-rank retrieved pages (Kleinberg, 1998). On the other hand, co-citation analysis[2] is used to detect networks of authors or to map topics and authors or journals (White, 2003; Zitt & Bassecoulard, 1994). CiteSeer[3] provides a reference to related documents from co-citation. Other digital libraries provide a cross reference to related documents. DBLP[4] for example provides a co-author index that gives access to the co-authors' publications or collaborative colleagues as does the ACM Portal.[5]

Digital libraries usually deliver results under the form of lists of related elements (list of related publications or authors) even though it has been shown that graphical interfaces play an important role in displaying the results of analysis to users (Chen, 2002; Geroimenko & Chen, 2002). In this context, graphs or networks are powerful techniques of visualization mainly because linking concepts or elements together is a very common mining technique. Another reason is that a network is easily understandable even by a naïve user.

When analyzing a word-wide phenomenon, such as scientific activities, mapping topics and countries or detecting core sets of countries or regions allows one to detect important trends. Geographic maps are the most intuitive way to describe and explain the spatial organization of a phenomenon that involves geographically referenced data, that is to say data which has locational references within its structure. Geographic maps as a way to communicate information to users by visualizing database-type information are powerful (Ahlberg & Shneiderman, 1994), but do not allow the user to explore relationships and interact directly with the data or analyze data that is not explicitly geographic (e.g. textual data).

[1] http://www.isinet.com/

[2] A co-citation can be extracted when two references appear in the same published paper.

[3] http://citeseer.ist.psu.edu/

[4] http://www.informatik.uni-trier.de/~ley/db/index.html

[5] http://portal.acm.org/

We argue that combining geographic maps, mining tools and other visualization modes interactively is much more powerful. In our approach, geographic maps are used to display the results of an analysis to the user. More importantly, at any step the user can interact with the visualizations to filter the information (changing their focus) and visualize the selected information using another visualization mode or mine this targeted information. This method implements a powerful and interactive discovery process. This process can be applied to any type of textual information even if heterogeneous in terms of format (structured documents such as Inspec©[6] documents, or semi-structured documents such as web documents). Our contribution is intended to help scientists and decision makes to easily have a large and interactive overview of specific domain knowledge related to research and technology. More precisely, it helps them to grasp strategic key factors—actors, actions, relationships and dynamics—of the domain. It can also help in the process of elaborating hypotheses and assessing them—e.g. by visualizing raw information that supports a hypothesis.

In this paper we present the model we use to represent the information in such a context. We also present the different agents of our system and the way they interact to implement a powerful knowledge discovery process. We illustrate the overall process using a case study based on scientific papers from SIGIR (Special Interest Group on Information Retrieval Conference). The paper is organized as follows: Section 2 presents related works. Section 3 describes the case study. Section 4 describes the way information is represented in order to be in a format that can be efficiently analyzed and an example from the case study is presented. Section 5 introduces different mining functions that are implemented in the platform with their usefulness being illustrated through the case study. It also presents the graphical view modules that users are provided with, also illustrated through our case study. Section 6 describes how the different agents communicate in order to discover additional advanced information. Finally, we conclude the paper and present potential future works.

2. Related works

The system we present and the underlying knowledge discovery process combines Data Mining (DM) and Geographic Information Systems (GIS) functionalities in order to analyze scientific domains from publications. GIS are computer systems to handle geographically referenced data, that is to say data that is described in terms of location (for example, spatial co-ordinates measured as latitude and longitude) or characteristics of spatial features (for example an address). DM on the other hand aims at discovering unknown information from data by applying mining functions such as classification, clustering, or detection of dependencies, etc. Combining these two approaches is important because of the quantity of spatial data available (Buttenfield, Gahegan, Miller, & Yuan, 2000).

Different software combines mining and GIS functionalities. GeoMiner (Han, Koperski, & Stefanovic, 1997) extends DBMiner (Han et al., 1997) to geo-referenced data. Initially, DBMiner mines databases using multi-dimensional analysis (Widom, 1995). Through a combination with MapInfo software, it is possible to mine data that includes a geographic dimension. The analysis can be done at different levels of detail (state, country, etc.). MacEachren, Wachowicz, Edsall, and Haug (1999) also investigated the integration

[6] Inspec© http://www.iee.org/Publish/Inspec/About/index.cfm

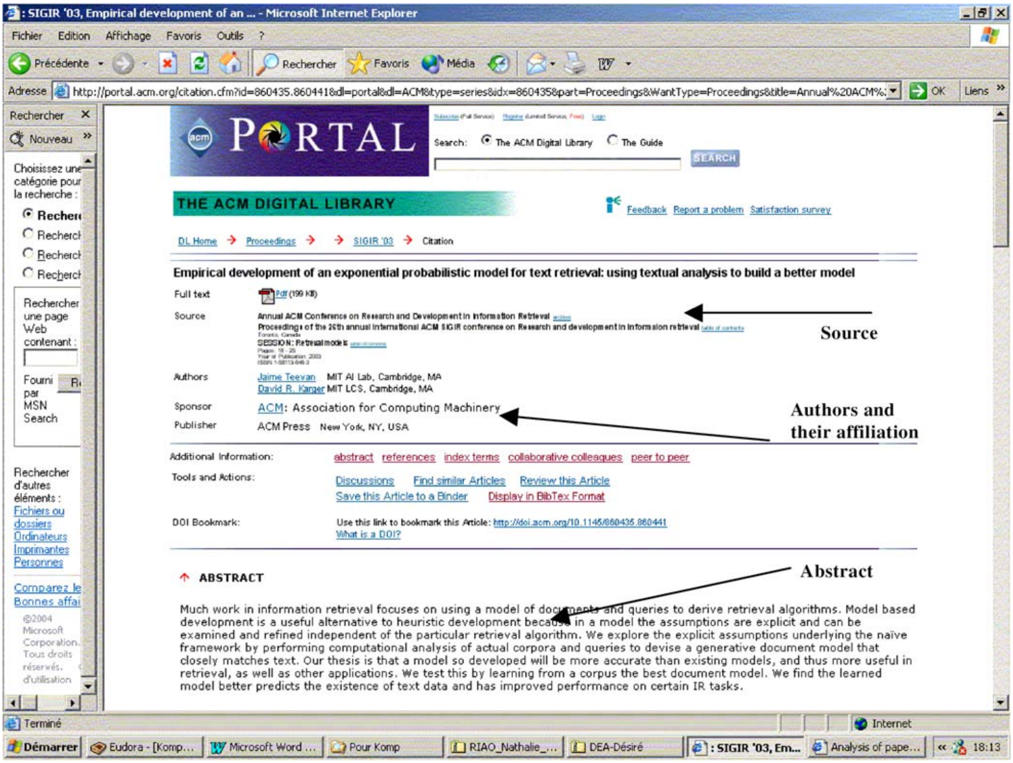
of data mining and geographic visualization in the context of spatiotemporal environmental data. Gahegan, Takatsuka, Wheeler, and Hardisty (2002) presents GeoVISTA studio, an environment designed for handling geospatial data. It can be used to build applications for geographic visualization that can combine different mining tools. In the same way, Descartes (Andrienko, Andrienko, & Gatalsky, 2000) is associated with the Kepler system to visualize statistical information such as election results, demographic data, etc. As with Spotfire (Ahlberg & Shneiderman, 1994) formerly known as IVEE, database content can be visualized using a variety of visualization tools.

Analyzing textual documents and their geographic dimension has been studied in different approaches. In Verbeek, Debackere, and Luwel (2003) the authors present an analysis of patents. They analyze the geographic distribution of science citation patterns including the flows between the regions. In Doré and Ojasoo (2001) publications from 48 countries in 18 disciplines over 12 years have been analyzed. In Mothe, Chrisment, Dousset, and Alaux (2003) a large set of documents is analyzed using an on-line analytical process considering different document dimensions, one of which being location. The methods used in these different approaches do not combine mining techniques and GIS functionalities. As a consequence the discovered information is not easy to visualize according to location. Visualization tools when analyzing a domain are a key issue. Boyack, Mane, and Börner (2004) reports a study on melanoma and the structure of the research on this topic from MedLine publications. Similarities between different element types, papers, genes, and proteins are calculated and the results are displayed under the form of networks using force-directed placement.

In Skupin (2004) and Skupin and Fabrikant (2003), the authors propose a method to map knowledge domains using self organized maps and use a nice metaphorical geographic language to depict to the studied domain. The input data are abstracts from scientific papers (abstracts submitted to the annual meeting of the Association of American Geographers). Unlike us the authors do not take advantage of the multi-faceted nature of the documents (for example titles, authors names, addresses, etc.). Also, apart from an interesting "semantic zoom" tool, the visualization model used is quite static. In our approach, geographic maps are more than a metaphor. They are used to display results of data mining that imply a geographic dimension. Moreover geographic maps are interactively changed according to messages delivered by other visualization modules or mining modules (implemented as agents). This feature is used to handle temporally geo-referenced data as well as to give focus to some target data.

3. Presentation of the case study and applications

We will illustrate the document mining process for domain analysis using a simple case study. The document set we chose should be considered only as a pedagogic illustration of the overall discovery principle. The system has been successfully used for competitive intelligence and science monitoring and evaluation purposes since 1991 in its earlier versions. Among real studies that have been conduced using the system, we can list: analysis of the TPV department (Plant Health and the Environment Department) of the French INRA institute (National Institute for Agricultural Research) (2002), an evaluation of INRA research in the domain of ecology, environment and engineering (2003), a study on OMEGA3 for Fabre laboratories and a study on the management of the pesticide Gaucho® crisis (2004). We carried out four health-related studies for the French Health



http://portal.acm.org/citation.cfm?id=860435.860441&dl=portal&dl=ACM&type=series&idx=860435&part=Proceedings&WantType=Proceedings&title=Annual%20ACM%

Choisissez une catégorie pour la recherche :
● Recher
○ Recherch
○ Recherch
○ Recherch

Rechercher une page Web contenant :
Fourni par MSN Search

Rechercher d'autres éléments :
Fichiers ou dossiers
Ordinateurs
Imprimantes
Personnes

Comparez le
Bonnes affai
©2004 Microsoft Corporation. Tous droits réservés. d'utilisation

Subscribe (Full Service) Register (Limited Service, Free) Login

Search: ● The ACM Digital Library ○ The Guide

THE ACM DIGITAL LIBRARY

Feedback Report a problem Satisfaction survey

DL Home → Proceedings → → SIGIR '03 → Citation

Empirical development of an exponential probabilistic model for text retrieval: using textual analysis to build a better model

Full text	Pdf (199 KB)
Source	Annual ACM Conference on Research and Development in Information Retrieval archive Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval table of contents Toronto, Canada SESSION: Retrieval models table of contents Pages: 18 - 25 Year of Publication: 2003 ISBN 1-58113-646-3
Authors	Jaime Teevan MIT AI Lab, Cambridge, MA David R. Karger MIT LCS, Cambridge, MA
Sponsor	ACM: Association for Computing Machinery
Publisher	ACM Press New York, NY, USA
Additional Information:	abstract references index terms collaborative colleagues peer to peer
Tools and Actions:	Discussions Find similar Articles Review this Article Save this Article to a Binder Display in BibTex Format
DOI Bookmark:	Use this link to bookmark this Article: http://doi.acm.org/10.1145/860435.860441 What is a DOI?

ABSTRACT

Much work in information retrieval focuses on using a model of documents and queries to derive retrieval algorithms. Model based development is a useful alternative to heuristic development because in a model the assumptions are explicit and can be examined and refined independent of the particular retrieval algorithm. We explore the explicit assumptions underlying the naïve framework by performing computational analysis of actual corpora and queries to devise a generative document model that closely matches text. Our thesis is that a model so developed will be more accurate than existing models, and thus more useful in retrieval, as well as other applications. We test this by learning from a corpus the best document model. We find the learned model better predicts the existence of text data and has improved performance on certain IR tasks.

Source

Authors and their affiliation

Abstract

Fig. 1. Harvesting documents from ACM Portal.

and Medical Research National Institute (apatite, apoptosis, flu, and epilepsy). For the French ministry of defense, we carried out analysis on the following topics: steel, information war, nuclear power, scientific literature on sub-marine detection. While the results of these studies are confidential, a simpler set of documents is sufficient to understand the power of the methodology used.

The set of documents we chose to illustrate the paper is composed of the long papers published in SIGIR conferences. This information has been extracted from the ACM Portal (see Fig. 1). In this paper we will use two document sets. One is composed of recent papers from 2000 to 2003 and consists in 173 papers in total. The second set is composed of the papers from 1991 to 2002 and consists of 440 papers.

Regarding geo-related information, the type of advanced information that can be discovered ranges from the main actors in terms of location of a domain (countries, laboratories or institutes), the specificities of a location (according to a domain, what are the sub-topics that characterize one location) and their evolution, links between location and dependencies including strength of the links regardless of either concrete collaboration (e.g. based on co-authoring) or without concrete collaboration (e.g. shared topics or specificities, shared behavior, etc.).

To make possible such analysis, a common representation of the information is adopted. The way we represent textual information and location is explained in the next section.

4. Information representation

4.1. Multi-dimensional representation

4.1.1. Document facets

Documents and texts express a vast and rich range of information that traditional information indexing does not take into account (Hearst, 1999). Indeed, usually in information retrieval applications, documents are viewed as bags of weighted words (Rijsbergen, 1979). That means that document indexing results in a multi-dimensional information space representation based on a single facet—the free-text content facet. As a result, the semantics of the information are diminished. To solve this problem, we promote an approach in which documents are represented in a multifaceted multi-dimensional space (Mothe et al., 2003). In this multi-dimensional representation, each facet corresponds to a point of view that may be of interest for the user with which a semantic is associated. Indeed, when considering a document as a bag of words, each term is considered in the same way. For example, the author of the document will be considered in the same way as any other term within the document. On the other hand, when considering multifaceted multi-dimensional representation as we suggest, each term is related to a context (facet) and thus carries more semantics. Facets can be viewed as "meta dimensions". In the rest of the paper the word dimension refers to facet.

One of the important dimensions corresponds to the geo-reference(s) associated with publications and corresponds to the producer(s) of the publication. This dimension is organized along a hierarchy of concepts (specificity/generality relationships). One level of the hierarchy is the institutes/organizations level. The upper levels are countries then continents. Institutes can be defined by their spatial co-ordinates whereas the other levels are areas defined by their borders. The levels of the hierarchy are set ones for all documents. For each document, its geo-reference information is extracted from the head of the document using the principle explained in Section 4.2. In the multi-dimensional representations we propose concept hierarchies related to each dimension can be handled. This is a more general approach than the idea of 'simple' concept hierarchies developed by Han and Fu (1994) and Han (1995). In fact there are two levels of concept hierarchies which are related to the following:

- Dimensions: each sub-concept defines a sub-dimension. For example "authors" can be split into two sub-dimensions which are "principal authors" and "collaborating authors". Another example is the "key words" dimension which can be split into "major key words" and "minor key words"—as in Medline the bibliographic database from the National Library of Medicine where descriptors are split into major and minor descriptors.
- Content: the content of a dimension can be organized along a concept hierarchy (this is the common approach).

Note that hierarchies of content can always induce hierarchies of dimension but the opposite is not true, as in the two examples given above. Hierarchies of concepts related to dimensions therefore provide a more general approach.

The geo-references associated with a document are not the only dimension we are interested in. Other document dimensions are "temporal references" (date of publication), "content" (which can be in turn subdivided into several points of view or sub-dimensions such as techniques used, names, locations, etc.). In some domains, pre-defined 'content' dimensions exist. In such a case, we simply use them. This is the case for example in medicine (Medical Subject Headings[7] Hierarchy) or in astronomy (IAU thesaurus[8]) and in many other domains. In cases where an ontology of the domain exists, we infer a hierarchy from it (Aussenac-Gilles & Mothe, 2004) before using it. In the other cases, where no hierarchy of concepts is associated, we simply extract the most discriminating terms using traditional indexing techniques (Frakes & Baeza-Yates, 1992). In the later case, the resulting dimension is flat. This is also considered as an explicit dimension of which "authors" is an example. Implicit dimensions are the ones generated from a dimension hierarchy of concepts.

Multi-dimensional representation has several advantages. Firstly, as we said before, each term corresponds to a context. For example, given an institute name in the 'producer of publication' hierarchy, it corresponds to the institute an author belongs to and not a name of an institute that is used in the content of a publication. Secondly, it is possible to associate documents to these hierarchies in an automatic way as is presented Section 4.2 (this can help answer questions such as "what are the institutes a given group of authors belongs to?" and "which countries are related to a given institute?"). Finally it is possible to browse and mine a collection of documents that is represented in this way.

4.1.2. Facets of the case study

A document as provided by ACM portal contains different types of information that can be extracted: the source, the title of the publication, its authors and affiliations, the ACM index terms and the publication abstract. According to this case study, different facets have been defined:

- Producer: at the most specific level this corresponds to the author names whereas at the most generic level it corresponds to the continent of the author's affiliation.
- Topic: corresponds to the concepts contained in the publications. Here we extract the topic from the title and the abstract. As a result this point of view is flat. However, we could have used the concept hierarchy provided by the ACM portal (index terms).
- Time: corresponds to the publication date. Because SIGIR is an annual conference, in that case the lowest level of granularity of the information is the year.

4.2. Mapping documents and generating points of view

Information extraction and document categorization techniques are used in order to map documents into each dimension (either explicitly or implicitly). After analysis, a document is represented by several sets of 'values' or 'dimension instantiations' that correspond to concepts from each dimension. For example, when considering the "producer of the publication" point of view, a document that has two institutions as "authors" will be associated with two nodes (instances) at the institute level—the hierarchy corresponding to that point of view. Implicitly, this document will also be associated to the parent nodes of these two institute-nodes (in our case the organizations they belong to).

[7] http://www.ulst.ac.uk/library/sci/MeSH.htm

[8] http://msowww.anu.edu.au/library/thesaurus/english/

4.2.1. Information extraction: mapping documents and hierarchies

According to our approach, a document is described according to a schema of extraction. The *schema of extraction* is the complete meta-description of the document set—that is to say all dimensions that can be used from the documents. The schema also describes the way the values of these attributes can be extracted according to the source. This schema has to be sufficiently formal to be source independent and yet sufficiently functional to be used in real applications to extract necessary dimension instantiations. Thus, the main purpose of the schema of extraction is to provide a description of how the useful information is to be extracted from a document set. The schema is fully defined by the five following components:

- The schema of extraction describes and uses the explicit *structure of the document set*. This defines the initial structure of documents that is to say meta-data that is marked-up in the document source.
- The schema of extraction defines the implicit *extraction structure:* This structure defines the list of useful dimensions that can be processed and their relationships to the explicit structure.
- The schema of extraction lists the *extraction rule set* which describes how these dimensions instances can be extracted according to the structure of the document set, (template filling out).
- The schema of extraction instantiates *semantic functions* for each dimension. These functions are used to solve semantic conflicts—such as synonymy—among the set of values of each dimension.
- The schema of extraction also instances *filtering functions*: these functions permit focusing on a reduced set of values when dealing with a given dimension—e.g. stop lists, important concept instances.

For more details of our information extraction model see Chrisment, Dkaki, Dousset, and Mothe (1997).

Note that this approach allows the handling of both structured and non-structured document sets. In the case of non-structured elements of information, we use categorization techniques in order to map documents and hierarchies. Concept hierarchies and documents can be associated in an intuitive way by considering each concept of the hierarchy as a category. Some approaches have been proposed in order to consider the hierarchical structure of the categories as opposed to a flat categorization (Weigend, Weiner, & Peterson, 1999). We developed our own method to associate a document with the concepts from the different concept hierarchies in which the hierarchical structure is taken into account (Mothe et al., 2003). In our model a document can be associated with several hierarchies and several concepts through the use of the semantic and filtering functions described above.

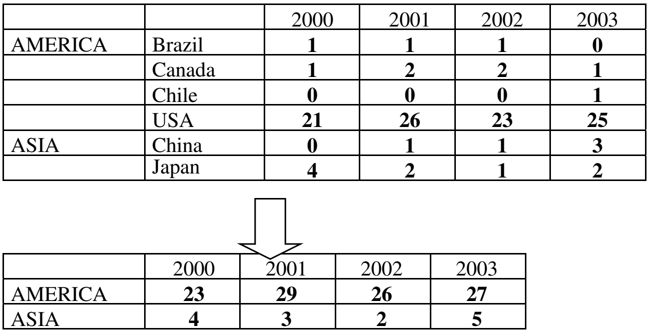
Table 1 indicates the number of different values for the document facets of the case study.

4.3. Information summarization

A document collection is summarized in the form of 2D matrices for which each matrix column or row corresponds to a dimension instantiations, filtered and reformulated using associated semantic and filtering functions. These matrices correspond to co-occurrence

Table 1
Number of different values for the document facets of the case study

# 1 document set (2000–2003)		# 2 document set (1991–2002)	
Facets	Number of different values	Facets	Number of different values
Authors	369	Authors	
Author's country	22	Author's country	31
Keywords	330	Keywords	
Dates of publication	4	Dates of publication	12



		2000	2001	2002	2003
AMERICA	Brazil	**1**	**1**	**1**	**0**
	Canada	**1**	**2**	**2**	**1**
	Chile	**0**	**0**	**0**	**1**
	USA	**21**	**26**	**23**	**25**
ASIA	China	**0**	**1**	**1**	**3**
	Japan	**4**	**2**	**1**	**2**

	2000	2001	2002	2003
AMERICA	**23**	**29**	**26**	**27**
ASIA	**4**	**3**	**2**	**5**

Fig. 2. Extract of a 2D-summarization structure using the producer (geo-reference) and temporal reference points of view from the case study.

matrices. In the specific case where different dimensions are considered, the matrix is a contingency table. Contingency tables are efficient in representing summarized information (Zembowicz & Zytkow, 1996). In addition, many data analysis methods use matrices or contingency tables as entries. Clustering or other methods can thus be applied as explained in Section 5.

Because the points of view are hierarchical, it is possible to consider the 2D tables at different levels of detail using roll-up and drill-down operators as defined in OLAP systems (Chrisment, Dousset, Karouach, & Mothe, 2003; Widom, 1995).

For example, Fig. 2 presents the results of crossing publication producers (lines) and years (columns) on the case study. The contingency table depicts the number of publications of each 'category'. A publication for which an author is from China is counted as a publication from ASIA.

5. Information mining and visualization agents

Different kinds of mining methods have been defined in the literature that can be applied on contingency tables or tables derived from them. We do not define new methods but rather lean on existing analysis methods. However our contribution is to implement them as agents that collaborate to provide synergy in the process of discovering information. Another contribution is to associate these functions with powerful visualization tools. As a result of users' demands some agents are activated—depending on their nature, they provide their results either to the user or to other agents as described Section 6.

In this section, we present only some of the functions, but other functions are implemented. Among these functions, we will present the spreadsheet functionalities and associated tools, the agglomerative classification and the network agent. We will also present the map visualization agent. For each of them, we will present one example of its usefulness when analyzing a set of documents.

The platform is also composed of different visualization tools: spreadsheet, histograms, graphs, 4D-views and geographic maps. These agents communicate with the mining agents and with each other, in this latter case mainly to compare result views.

5.1. Spreadsheet agent

5.1.1. Functionalities

The spreadsheet agent allows one to manipulate directly the 2D tables. This agent offers the user the main spreadsheet functionalities (deleting, adding columns and rows, reordering of rows and columns according to the cells' values, operation on the values such as mean, additions, histograms, etc.).

This agent is used in order:

- To obtain normalized data when the initial data is not e.g. dividing the values by the sum. For example, from a table that cross authors-rows and dates-columns, dividing the row-values by the sum of these values allows one to get the relative contributions of the authors.
- To detect the most frequent items (Fig. 3a).
- To detect evolution in data, for example, from a table that crosses countries and years a histogram on a specific country allows one to visualize the evolution of the publication activity of that country (Fig. 3b–d).

5.1.2. Spreadsheet to discover evolution of the countries contribution

In this example, we consider the second set of documents (from 1991 to 2002) for analysis and the data from 2003 to evaluate the predictions our system makes.

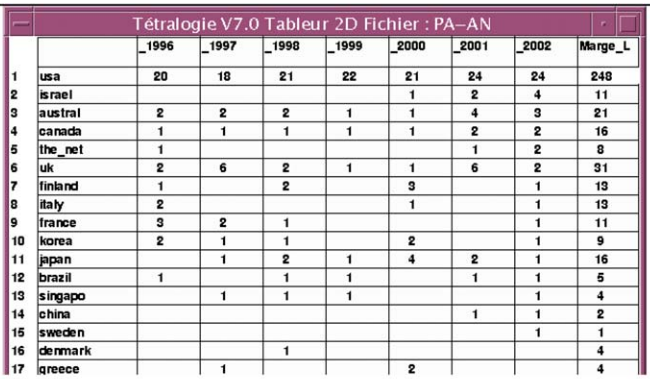
The system provides a table that crosses countries and dates. In this sample of documents, 12 years and 31 countries occur at least once. A paper that is co-authored by people from two different countries will be counted twice.

5.2. Classification methods

2D summarization structures as defined in Section 4.3 fit well *classification* methods, it is possible to group together items (lines) according to characters (columns). The platform includes two types of classification agents: Agglomerative Hierarchical Clustering (AHC) and Classification by Partition (CbyP).

5.2.1. Agglomerative hierarchical clustering

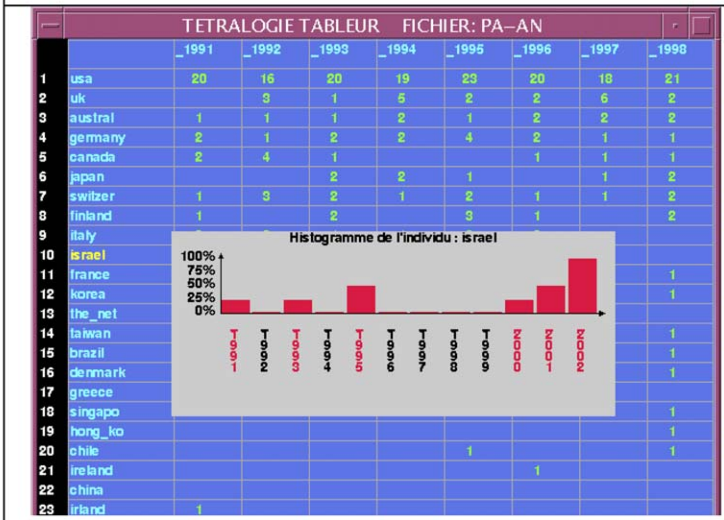
AHC is used to classify the elements according to their similarities. Two classes are grouped together if they are the closest couple of classes. At the beginning of the process, each element corresponds to a class. The process ends when the target number of classes is reached.



		_1996	_1997	_1998	_1999	_2000	_2001	_2002	Marge_L
1	usa	20	18	21	22	21	24	24	248
2	israel					1	2	4	11
3	austral	2	2	2	1	1	4	3	21
4	canada	1	1	1	1	1	2	2	16
5	the_net	1					1	2	8
6	uk	2	6	2	1	1	6	2	31
7	finland	1		2		3		1	13
8	italy	2				1		1	13
9	france	3	2	1				1	11
10	korea	2	1	1		2		1	9
11	japan		1	2	1	4	2	1	16
12	brazil	1		1	1		1	1	5
13	singapo		1	1	1			1	4
14	china						1	1	2
15	sweden							1	1
16	denmark			1					4
17	greece		1			2			4

(a) Number of paper per countries and per year

Sorting the columns in the chronological order and then the raw according to the cell values for a given date (here 2002), it is possible to see the most 'prolific' countries, the ones that contribute the most, depending on the year. Additionally, in this example a column has been added on user's demand that contains the total number of item per row (number of publication per year).



TETRALOGIE TABLEUR FICHIER: PA–AN
Histogramme de l'individu : israel
100%
75%
50%
25%
0%
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002

		_1991	_1992	_1993	_1994	_1995	_1996	_1997	_1998
1	usa	20	16	20	19	23	20	18	21
2	uk		3	1	5	2	2	6	2
3	austral	1	1	1	2	1	2	2	2
4	germany	2	1	2	2	4	2	1	1
5	canada	2	4	1			1	1	1
6	japan			2	2	1		1	2
7	switzer	1	3	2	1	2	1	1	2
8	finland	1		2		3	1		2
9	italy								
10	israel								
11	france								1
12	korea								1
13	the_net								
14	taiwan								1
15	brazil								1
16	denmark								1
17	greece								
18	singapo								1
19	hong_ko								1
20	chile					1			1
21	ireland						1		
22	china								
23	irland	1							

(b) Evolution of the publication number for Israel

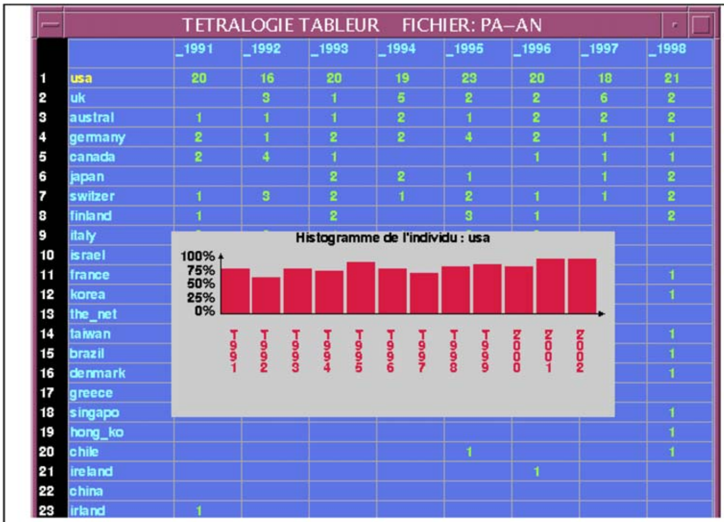
From the same matrix, a histogram for Israel shows the publication evolution. From this, it is possible to predict Israel will publish in future years. Indeed, Israel published 1 paper in 2003 and 3 papers in 2004.

Fig. 3. Crossing countries and years to predict country contribution to SIGIR.

This mining function is useful when one wants to detect elements that follow the same behavior. For example, it can be used to group together laboratories that have the same topics of interest (working on the same areas). In that case the source table would cross topics and laboratories.

5.2.2. Classification by partition

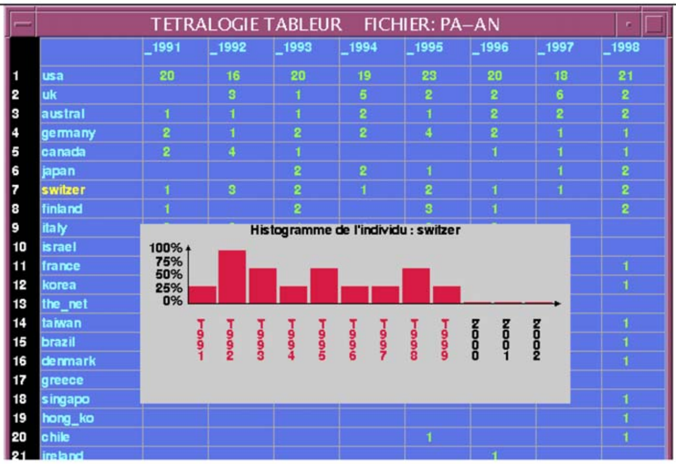
A CbyP is a non-hierarchical and supervised classification used to classify data according to predefined classes. It is also used as a technique of description and analysis—generally in partnership with factor analyses—as well as a reduction technique as the set


TETRALOGIE TABLEUR FICHIER: PA–AN
Histogramme de l'individu : usa
100%
75%
50%
25%
0%
T991
T992
T993
T994
T995
T996
T997
T998
T999
Z000
Z001
Z002

		_1991	_1992	_1993	_1994	_1995	_1996	_1997	_1998
1	usa	20	16	20	19	23	20	18	21
2	uk		3	1	5	2	2	6	2
3	austral	1	1	1	2	1	2	2	2
4	germany	2	1	2	2	4	2	1	1
5	canada	2	4	1			1	1	1
6	japan			2	2	1		1	2
7	switzer	1	3	2	1	2	1	1	2
8	finland	1		2		3	1		2
9	italy	[illegible]	[illegible]			[illegible]	[illegible]		
10	israel								
11	france								1
12	korea								1
13	the_net								
14	taiwan								1
15	brazil								1
16	denmark								1
17	greece								
18	singapo								1
19	hong_ko								1
20	chile					1			1
21	ireland						1		
22	china								
23	irland	1							

(c) Evolution of the publication number for USA

Regarding the contribution of the USA, the system indicates their level of publication is stable whatever the year. It is possible to predict this will continue. Indeed USA published 26, 23 and 25 papers respectively in 2001, 2002 and 2003.


TETRALOGIE TABLEUR FICHIER: PA–AN
Histogramme de l'individu : switzer
100%
75%
50%
25%
0%
T991
T992
T993
T994
T995
T996
T997
T998
T999
Z000
Z001
Z002

		_1991	_1992	_1993	_1994	_1995	_1996	_1997	_1998
1	usa	20	16	20	19	23	20	18	21
2	uk		3	1	5	2	2	6	2
3	austral	1	1	1	2	1	2	2	2
4	germany	2	1	2	2	4	2	1	1
5	canada	2	4	1			1	1	1
6	japan			2	2	1		1	2
7	switzer	1	3	2	1	2	1	1	2
8	finland	1		2		3	1		2
9	italy	[illegible]	[illegible]			[illegible]	[illegible]		
10	israel								
11	france								1
12	korea								1
13	the_net								
14	taiwan								1
15	brazil								1
16	denmark								1
17	greece								
18	singapo								1
19	hong_ko								1
20	chile					1			1
21	ireland						1		

(d) Evolution of the publication number for Switzerland

Finally, considering Switzerland, the system predicts the contribution to SIGIR of this country is decreasing. There is no publication in 2003.

Fig. 3 (*continued*)

centroids of the classes can be considered as representative of the entire data set. The initial method can be principally credited to Forgy (1965). The general principle of the method is aggregation around mobile centroids. This makes CbyP similar to, but not identical with, many classification techniques such as "dynamic clouds" (Diday, 1974) and *k*-means introduced by MacQueen (1967). In fact CbyP is some times labeled as "batch" *k*-means.

This function can be used when there is a target group number to obtain. When the user wants to partition the laboratories into four groups regardless of the topics they are working on, a CbyP using four centroids would be the most appropriate.

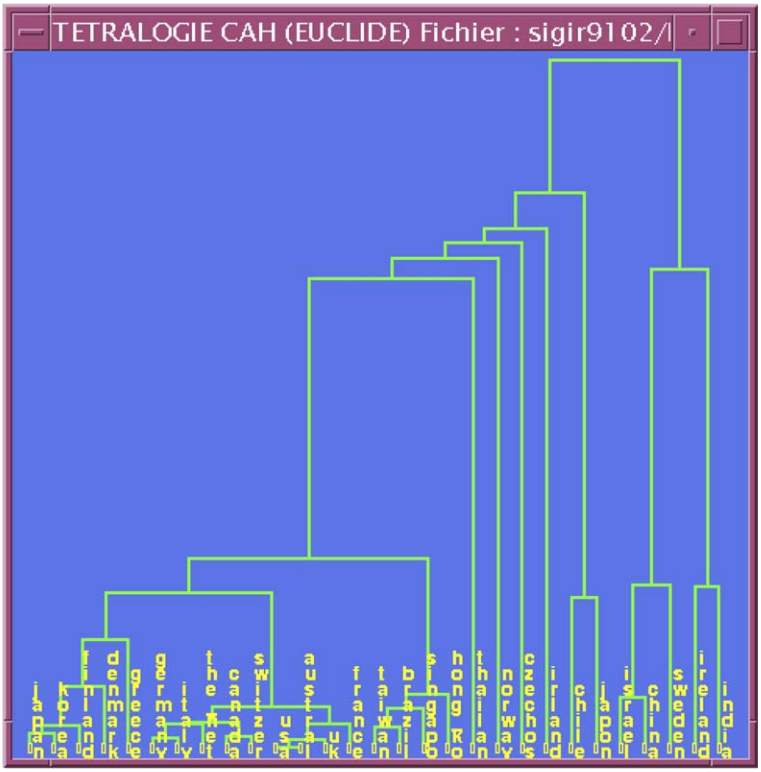


Fig. 4. Grouping countries according to their publishing profiles.

5.2.3. Clustering countries according to their contribution profiles

In this example, we considered again the same matrix. First each year and each country has been normalized so that we obtain their profile rather than their contribution in terms of number. For example, the USA is publishing regularly (comparable number of paper every year). A country that did the same, even if only publishing one paper per year would have the same profile. Fig. 4 displays the results. Australia is the closest to America regarding its profile.

5.3. Geographic maps

5.3.1. Principle

Geographic maps are used when geo-referenced data is handled. Geographic map agents use a 2D matrix in which one of the dimensions corresponds to geo-referenced data (Karouach, 2003). As we stated previously, institutes are geo-referenced data that can be defined by their spatial co-ordinates. However, in the current module, the minimum granularity of representation is at the level of country. However, using specific databases that provide spatial co-ordinates, it would be possible to change increase this granularity.

To illustrate the usefulness of geographic maps in our approach, we show how they can be used in order to visualize the contribution of each country to the document set—we also

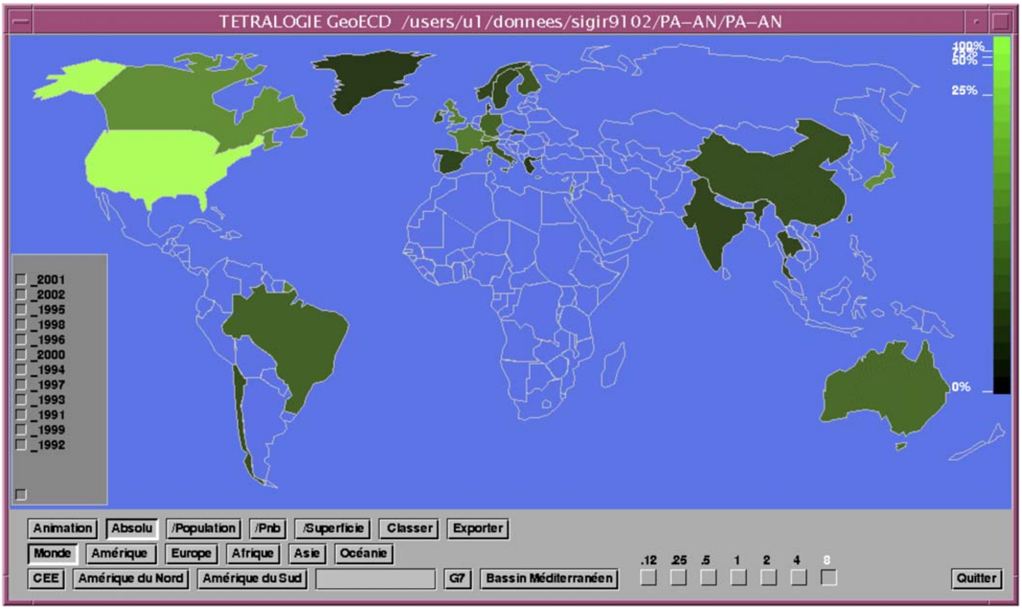


Fig. 5. Contribution of the different countries along the entire period (1991–2002).

show how evolution can be taken into account in this analysis. This analysis is based on authorship.

5.3.2. Maps to visualize the contribution of each countries

The previous section explains how an analysis can make clearer the specificity of some countries in term of topics. The relative contribution of each country is another type of information that explains how a domain is structured. Regarding our case study, it corresponds to the contribution in term of number of publications that have been accepted at the SIGIR conference. This contribution can be analyzed globally or year by year in order to understand the evolution.

The starting point of this representation is a 2D matrix in which lines correspond to countries whereas columns correspond to the year of publication. The matrix indicates the number of publications that are written by an author from a given country for a given year. Note that publications that have two co-authors from a single country are counted only once whereas a publication that is co-authored by authors from two different countries is counted twice (one for each country).

In Fig. 5, 4 years of publication are considered. The greener[9] (brighter) a country is displayed, the more publications it has. The scale of coloring is not linear because otherwise the only country that would have been colored would be the USA. Instead, we applied a non-linear scale so that contributing countries are colored, even if their contribution is comparatively less. Countries that are not colored made no contribution in the analyzed set of documents.

[9] For interpretation of the references in colour in figures, the reader is referred to the web version of this article.

More importantly, it can be interesting to visualize the evolution of the contribution of the countries through time. In Fig. 5, the country contributions are considered whatever the year of publication is, but it is also possible to visualize only one year (or selected years). In Fig. 6a for example, only 1991 is displayed; in the same way Fig. 6b displays the information for 2002. In Fig. 6, as in the Fig. 5, non-colored countries have no contribution during the whole period (e.g. African countries). Countries in green are the countries that contribute in the considered year and countries in red are countries that do not contribute in the considered year even if they contribute in the studied period (e.g. China contributes during the period; however this country did not contribute in 1991 but does in 2002).

From Figs. 5 and 6 other interesting knowledge can be extracted: UK proportionally contributes more over the entire period than just in 2000 (the brightness of the green gives this information). This is the opposite for Finland which contributes more in 2000 than during the entire period (proportionally). In 2002, Finland organized SIGIR.

This type of analysis is useful to show the emerging countries in a domain. When combined with the previous analysis on correlation between topics and countries, the added value of such analysis is important.

5.4. Graphs

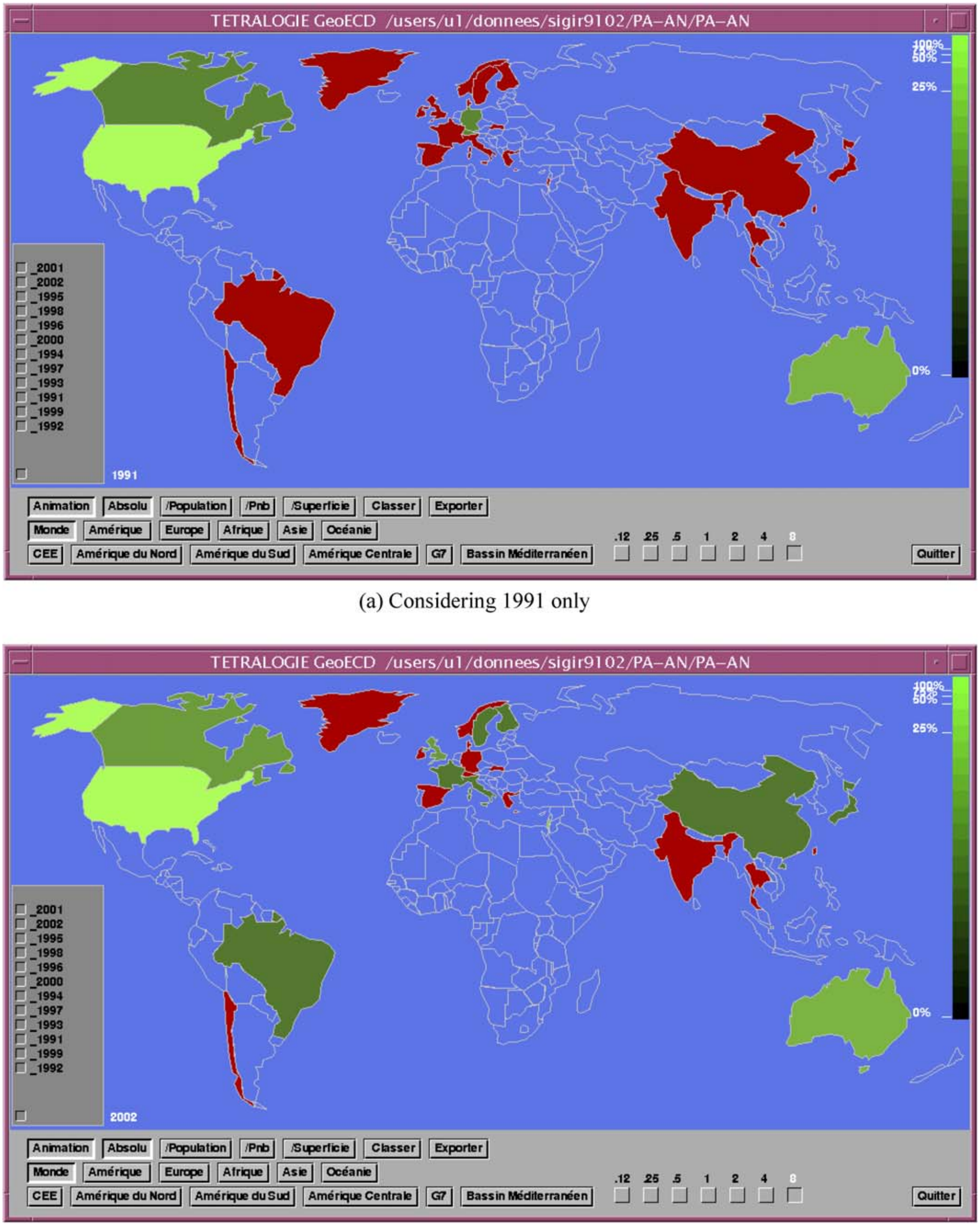
5.4.1. Principle

Graphs are among the visualization tools most commonly used in the literature, as linking concepts or objects is the most common mining technique. Graph agents use 2D matrices of any type resulting from the pre-treatment of the raw information and corresponding to co-occurrence matrix. A 2D matrix can also result from a mining method. Graph nodes correspond to the values of the crossed items whereas edges reflect the strength of the co-occurrence value. Graph drawing is based on (Fruchterman & Reingold, 1991). In this type of algorithm, a graph node is considered as an object while an edge is considered as a spring. Edge weights correspond to either repulsion or attraction forces between the objects that in turn make them move in space. This keeps the vertices moving in the visualization space until an equilibrium position is reached. Once stabilized the spring system provides the best graph drawing or node placement.

We also considered the analysis of graph properties (Karouach, 2003; Truong, 2004) such as cluster analysis for which we use spectral analysis among other methods—some of them have been discussed above. To provide ways of identifying the most influential objects—authors, countries, etc., we provide centrality analysis methods such as degree, betweeness, or proximity analysis. We are now exploring methods to conduct structural analysis of graphs like in (Kleinberg, 1999) regarding hub/authority measures. To illustrate the usefulness of graphs in domain analysis as we promote, we present in the two next sections illustrations of how graphs are used to visualize collaborations at the country level (considering co-authoring) and common topics of interest of different countries (according to publication content).

5.4.2. Graphs to visualize collaboration between countries

Collaboration between countries is a source for technological activity and creativity around the world (Verbeek et al., 2003). Thus it is important to be able to know the countries that collaborate with each other and the strength of these collaborations



(a) Considering 1991 only

(b) Considering 2002 only

Fig. 6. Contribution of the different countries for one selected year.

(frequency). To analyze this type of collaboration, the starting point is a 2D matrix based on the *producer* point of view. The same attribute is considered, but at two levels of detail: country and author levels. That is to say lines correspond to authors' names

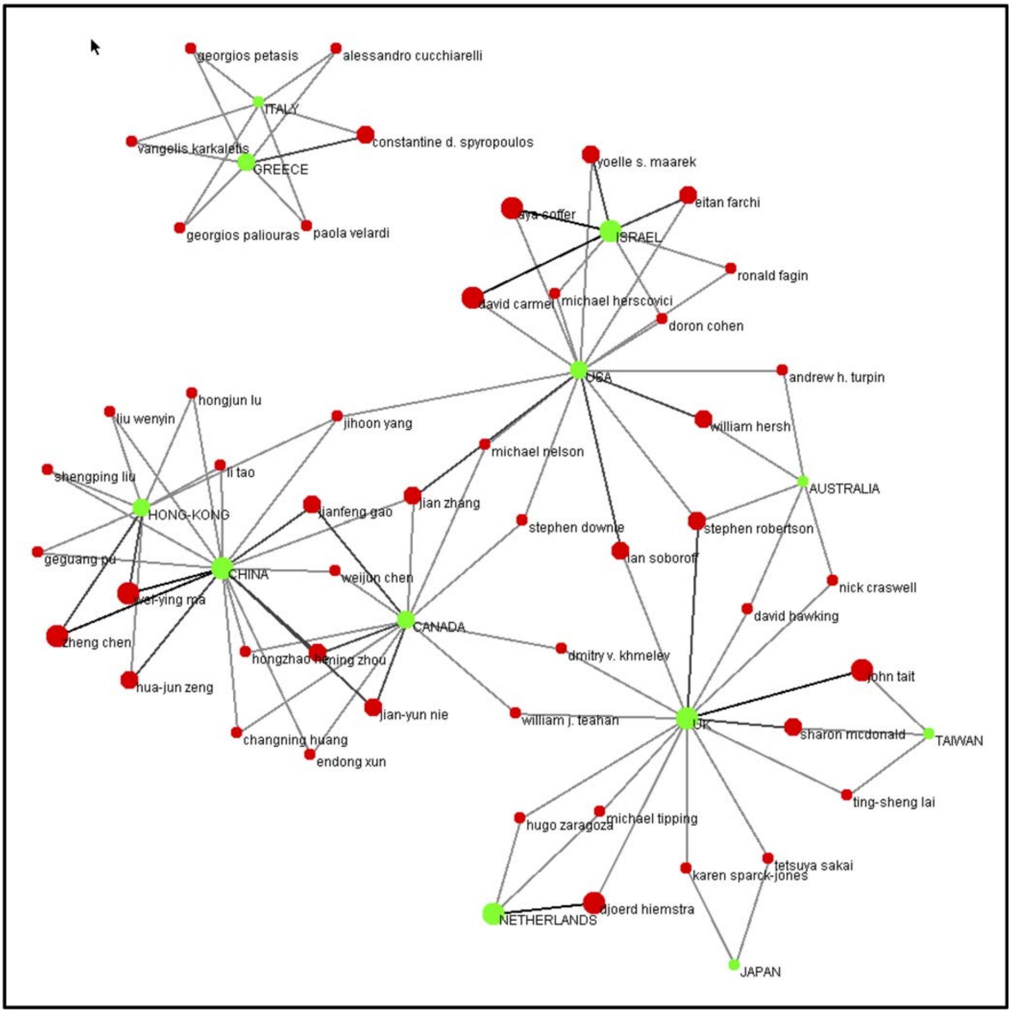


Fig. 7. Author/country network.

whereas columns correspond to countries. At the crossing of a line and a column, we obtain the number of publications that for a given author co-occur with a given country name (the country s/he belongs to or the country which a co-author belongs too). Note that in the case of multiple co-authors from the same country for a single publication, the co-occurrence is counted only once. We use the first document set for this analysis (2000–2003). In Fig. 7, countries appear in green whereas authors are displayed in red. Countries that are not correlated to other countries do not appear in this graph. That means that we only consider the publications that have at least two authors belonging to two different countries. The edges correspond to links that have been inferred between countries and authors.

Using this type of network representation, cooperation between countries appears in a single shot. For example, strong relationships are shown between China and Hong Kong and between Israel and USA. China and Hong Kong are not surprising considering the political point of view. Israel and USA relationships in IR are explained by the fact that a laboratory of the IBM Company is situated in Haifa (Israel) and publications are co-authored with IBM US (this can be validated when going back to the publication themselves). The power of this representation is that links are drawn, but more importantly, the explanation of the link can be seen. When considering the Netherlands and the UK for example, the association is manly due to 'Djoerd Hiemstra'. In the same way, the association between China and Canada is due to two persons: Jian Yun Nie (Canada) and Ming Zhou (China). The former author is from China, did his Ph.D. in France and has now a position at the University of Montreal, Canada. Another important link can be shown between the UK and Taiwan. The authors the link is due to are also shown (one is John Tait). This link also has a possible explanation. The latter author has had strong links with Taiwan for a while: in 1994 a Ph.D. student from Taiwan passed his Ph.D. in John Tait's group and is now in Taiwan; authors from Taiwan and from UK co-publish since at least 1998. This information is not in the document sample that is analyzed, but came from validation from the Web and from digital Libraries.

5.4.3. Graphs to extract the main topics of interest of the countries

All countries may not evenly contribute to a domain. Some countries may be more specialists in some sub-fields (for historical reasons or because the countries funding agencies support this sub-field for example). It is useful to know if some countries have specific

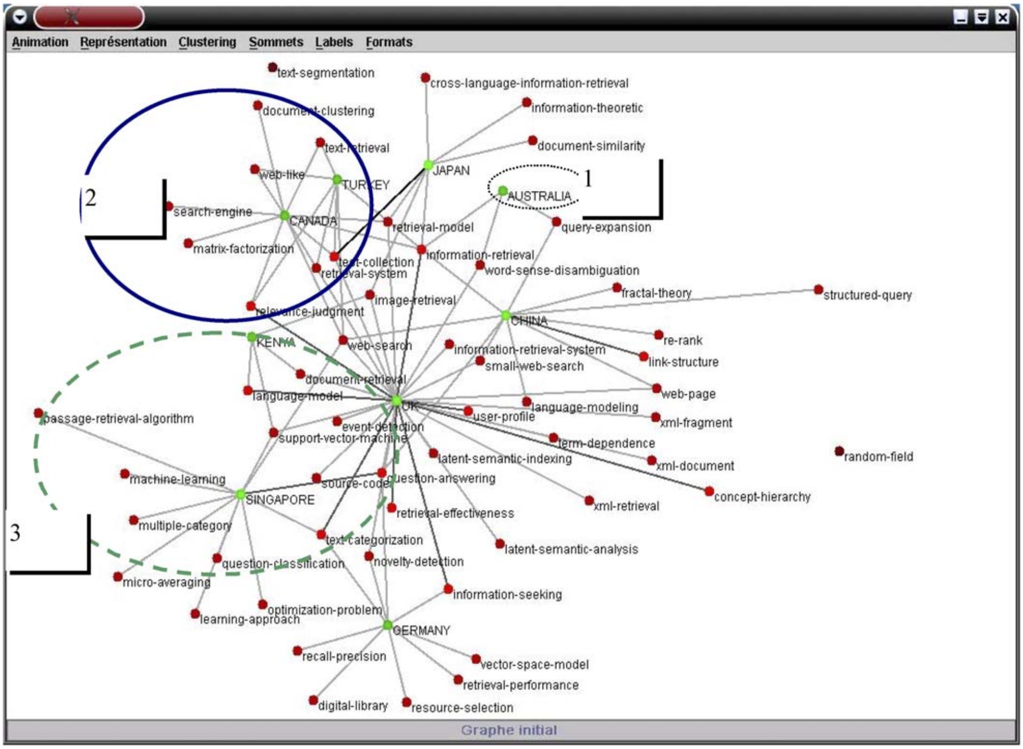


Fig. 8. Topics/country network.

interests and what are the topics that are common to all the countries. This is the purpose of this part of the analysis.

In the case of Fig. 8, the 2D matrix consists of lines that correspond to key words and columns that correspond to countries. Some interesting sub-networks have been circled on the figure. For example, Canada and Turkey are linked through common topics of interest and this link is not due to some publications that have been written by an author from Canada and another from Turkey (in Fig. 7 these two countries are not linked).

6. Communication between mining/visualization agents

In our approach, collaboration between agents is used in two ways. When a *single user* is involved one visualization agent can use the results of a mining agent. As an example the result of a classification based on geo-reference (the objects classified are countries) can be displayed on a geographic map. In that case, a different color will be associated with each class. In the same way, two visualization agents or two mining agents can communicate. This mechanism improves the mining process.

When several users are involved, collaboration between agents gives a way of collaboration/synergy between users. In this case, all the users analyze the same data at the same time over the network and need to share information or exchange points of views.

6.1. Communication principle

We provide several ways of agent communication based on inter process communication. Among those ways of communication we can mention:

1. Broadcast: queries are broadcast to agents and sometimes filtered by specialty. This kind of communication is used for example to focus on the same object in different visualization tools.
2. Publish/subscribe: agents provide services—mainly message broadcasts—which are used by agents to those services. For example, classification agents provide partitions to visualization agents. Another example is 4D visualization agents that 'export' their point of view—focus, rotations and zoom—to other 4D agents who explicitly register this kind of information. This make it possible for users to control each others 4D visualization of compatible data sets.

6.2. System architecture and agent communications

Fig. 9 illustrates the communication channels between mining and visualizing agents, whilst the architecture of the whole system is presented in Fig. 10.

As shown in Fig. 9, when activated, an agent registers itself with the communication facilitator. This enables the facilitator to have a comprehensive overview of active agents, their state (willingness to communicate), and the data they are handling. The facilitator also manages a database of raw data, numerical results, users' captured views and it analyses annotations.

Every message that is sent to the facilitator is forwarded to active and compatible agents which are dealing with data in relation to the sent data.

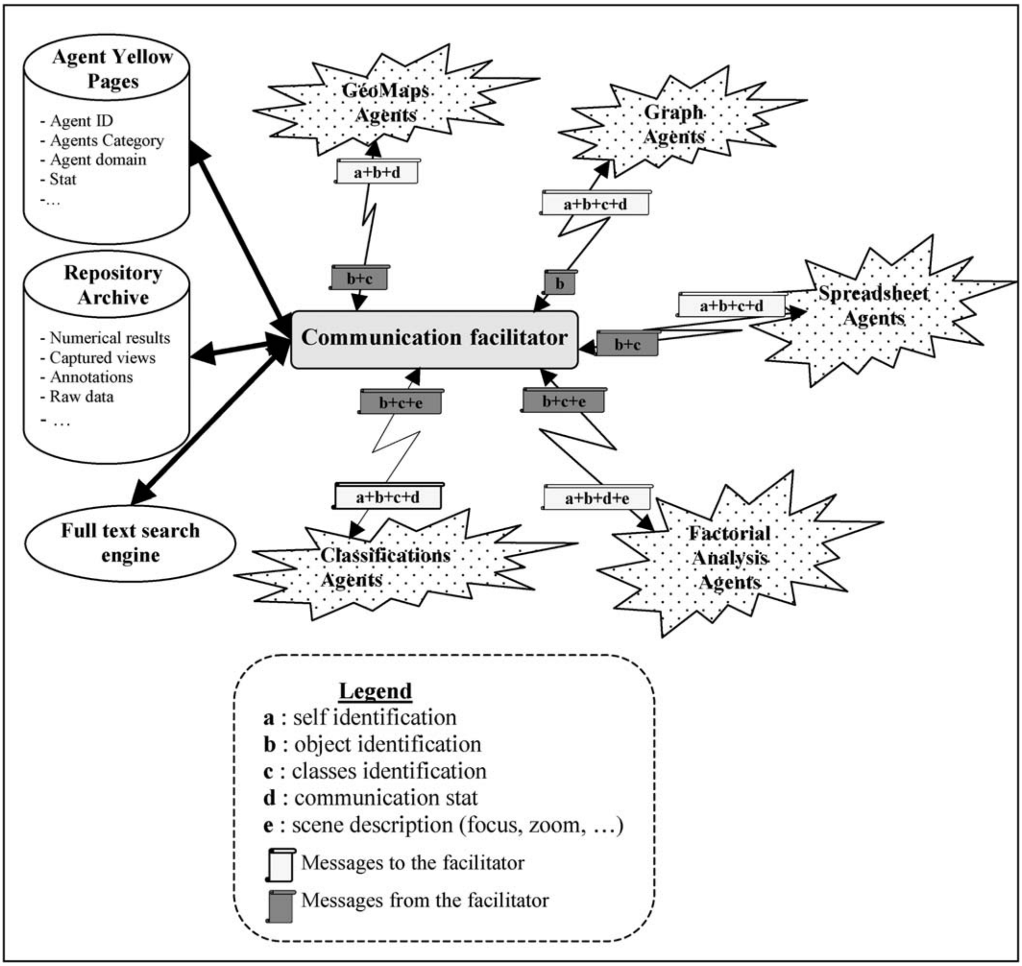


Fig. 9. Communication between agents.

Moreover, every agent has access through the facilitator to a search engine application to provide the user with raw information about analyzed or visualized information.

Any visualization modules can directly visualize summarized information or it can be mined. The results of the mining modules can also be visualized using the visualization modules. Mining and visualization modules can be combined in different ways (see Section 6.3).

6.3. Illustration of communication to discover information

In the previous section we indicate the importance of combining different types of analysis (e.g. contribution of the countries and correlation between countries and topics). Combining mining and visualizing tools has many applications. For example, combining

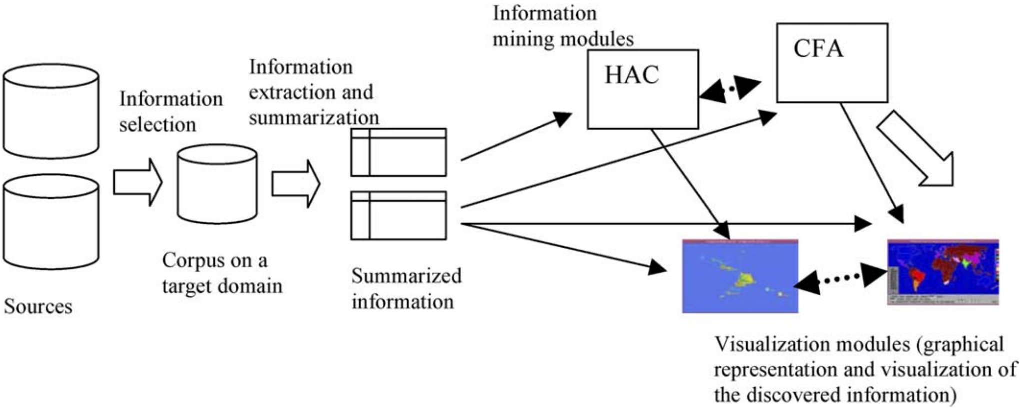

Information mining modules
HAC
CFA
Information selection
Information extraction and summarization
Sources
Corpus on a target domain
Summarized information
Visualization modules (graphical representation and visualization of the discovered information)

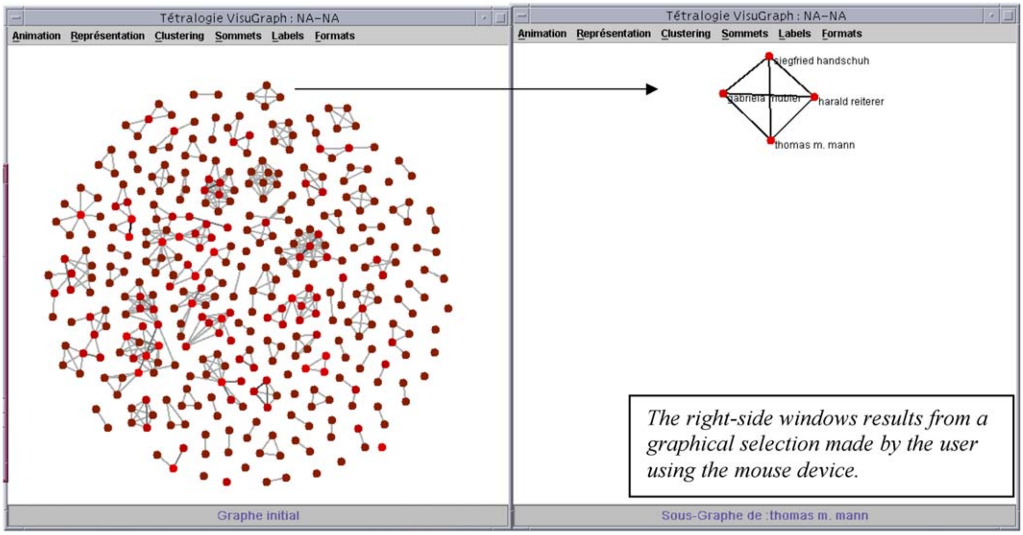
Fig. 10. Mining and visualization agents in the system architecture.

graphs (e.g. networks of authors) and maps can help understanding a phenomenon (countries that are close one to another, countries that share a common language, etc.). It is also possible to combine HAC and maps to visualize countries with the same publishing profiles.

In the rest of this section we depict the first example of combination: graphs and maps.

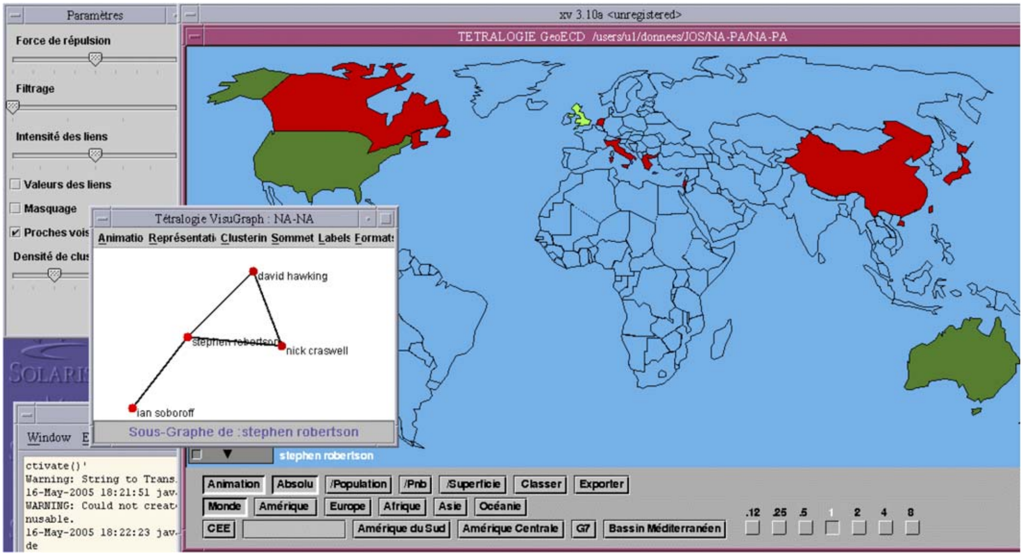
6.3.1. Combining networks and maps to visualize the collaborations of an author at a country level

We illustrate an efficient way to combine graphs and maps in order to understand better the location influence on author clusters.



The right-side windows results from a graphical selection made by the user using the mouse device.

Fig. 11. Co-authoring network.



```
ctivate()'
Warning: String to Trans.
16-May-2005 18:21:51 jav.
WARNING: Could not creat.
nusable.
16-May-2005 18:22:23 jav.
de
```

Fig. 12. Combining graphs and maps.

Fig. 11 depicts the co-authoring associations. Each node corresponds to an author whereas edges correspond to co-authoring association. Nodes are colored to show the importance in terms of contribution of the author (number of publications in the document set). The brighter a node is, the more the corresponding author contributes. In the screen shot, we removed author names to make the network more readable, but it is also possible to visualize the detail of each node. One can see that clusters of authors are clearly visible, showing that there is not a lot of co-authoring. It is possible to select a sub-graph on the left-side sub-window and visualize it on the right side window.

From a graph (right-side window, Fig. 11), it is also possible to select an author. In that case, the sub-graph corresponding to this selected node is automatically extracted: it includes the nodes that have at least one connection with the selected node (it is possible to choose the level of the hierarchy in order to include the direct neighbors or the neighbors of the neighbors). Fig. 12 shows the combination of the graph module and the map module. The country of the selected author appears in bright green (UK), the countries of the correlated authors appear in darker green. Again, the countries in orange are the countries that contribute to the document collection but in this case that are not in the selection (none of the authors selected are from these countries).

7. Conclusion

In this paper, we present a method to aid understanding of the state-of-the-art and the evolution of scientific communities and research topics. It is based on the analysis of document sets (scientific publications) from which information is extracted and mined. Our contribution is not to present new mining methods but rather to lean on existing ones. The platform we developed integrates different data analysis methods from the literature as well as different visualizing modules. After interesting attribute values are extracted from the documents, 2D tables are built that correspond to summarized data. 2D matrixes correspond to the input and output of the mining modules and input of the visualization modules. This makes it possible to make the modules communicate.

We illustrate our approach through a case study. Using this case study, it was not possible to present all the possibilities and combination our platform allows in order to analyze a domain (e.g. factorial analysis has not been illustrated) but we show interesting types of analysis that can be conduced using our methodology and our system. We focus on geo-referenced data and show different types of advanced information that can be extracted from documents. We also show how geo-references can be combined with other metadata to provide global views on a domain. Finally, we show that maps can be combined with other visualizations to provide the user with different views of the data.

Future work will concern the level of granularity the geographic map module takes into account. Currently the smallest level is the country, which is acceptable for some applications but not for all applications. Generally, a more detailed visualization is needed. Many GIS databases provide detailed information that could be included in our modules (both in the information extraction module and in the visualization module). A long term issue regarding this type of system is to define criteria of evaluation and benchmark collections as Information Retrieval community does with programs such as Text Retrieval Conference (trec.nist.gov).

References

Ahlberg, C., & Shneiderman, B. (1994). Visual information seeking: Tight coupling of dynamic query filters with starfield displays. In *CHI'94: Proceedings of the SIGCHI conference on Human factors in computing systems, Boston, Massachusetts, United States*, 0-89791-650-6 (pp. 313–317). New York, NY, USA: ACM Press.

Andrienko, N., Andrienko, G., & Gatalsky, P. (2000). Towards exploratory visualisation of spatio-temporal data. In *Proceedings of 3rd AGILE conference on geographic information science, Helsinki/Espoo*, pp. 137–142.

Aussenac-Gilles, N., & Mothe, J. (2004). Ontologies as background knowledge to explore document collections. In *Proceedings of international conference on computer assisted information retrieval*, pp. 129–142.

Boyack, K. W., Mane, K. K., & Börner, K. (2004). Mapping medline papers, genes, and proteins related to melanoma research. In *Proceedings of IV2004 conference, London, UK*, pp. 965–971.

Buttenfield, B., Gahegan, M., Miller, H. J., & Yuan, M. (2000). Geospatial datamining and knowledge discovery. University Consortium For Geographic Information Science Research, White paper. Available from http://www.ucgis.org/priorities/research/research_white/2000%20Papers/emerging/gkd.pdf.

Buter, R. K., & Noyons, E. C. M. (2002). Using bibliometric maps to visualise term distribution in scientific papers. In *Proceedings of sixth international conference on information visualisation*, p. 697.

Chen, C. (2002). Visualisation of knowledge structures. *Handbook of software engineering and knowledge engineering*, 2002.

Chrisment, C., Dkaki, T., Dousset, B., & Mothe, J. (1997). Extraction et synthèse de connaissances à partir de données hétérogènes. *Proceedings of Ingénierie des Systèmes d'Information, 5*(3), 367–400.

Chrisment, C., Dousset, B., Karouach, S., & Mothe, J. (2004). Information mining: extracting, exploring and visualising geo-referenced information. In *Proceedings of workshop on geographic information retrieval*, SIGIR 2004.

Diday, E. (1974). Recent progress in distance and similarity measures in pattern recognition. In *Second international joint conference on pattern recognition*, pp. 534–539.

Doré, J. C., & Ojasoo, T. (2001). How to analyze publication time trends by correspondence factor analysis: Analysis of publications by 48 countries in 18 disciplines over 12 years. *Journal of the American Society for Information Science and Technology, 52*(9), 763–769.

Forgy, E. (1965). Cluster analysis of multivariate data: Efficiency vs. interpretanility of classifications. *Biometrics, 21.*

Frakes, W., & Baeza-Yates, R. (Eds.). (1992). *Information retrieval: Algorithms and data structures.* Prentice-Hall.

Fruchterman, T., & Reingold, E. (1991). Graph drawing by force-direct placement. *Software Practice and Experience, 21*, 1129–1164.

Han, J. (1995). Mining knowledge at multiple concept levels. In *Proceedings of the 4th international conference on information and knowledge management (CIKM'95), Baltimore, Maryland*, pp. 19–24.

Han, J., & Fu, Y. (1994). Dynamic generation and refinement of concept hierarchies for knowledge discovery in databases. In *AAAI'94 workshop on knowledge discovery in databases (KDD'94), Seattle, WA*, pp. 157–168.

Han, J., Koperski, K., & Stefanovic, N. (1997). GeoMiner: a system prototype for spatial data mining. In *Proceedings of ACM–SIGMOD international conference on management*, pp. 553–556.

Hearst, M. A. (1999). Untangling text data mining. In *Proceedings of annual meeting of the association for computational linguistics.* Available from http://www.sims.berkeley.edu/~hearst.

Karouach, S. (2003). Visualisations interactives pour la découverte de connaissances: concepts, méthodes et outils, Ph.D. Thesis.

Kleinberg, J. M. (1998). Authoritative sources in a hyperlinked environment. In *Proceedings of ACM–SIAM symposium on discrete algorithms*, pp. 668–677. Available from http://www.cs.cornell.edu/home/kleinber/auth.ps.

Kleinberg, J. (1999). Hubs, authorities, and communities. *ACM Computing Surveys, 31*(4).

Gahegan, M., Takatsuka, M., Wheeler, M., & Hardisty, F. (2002). Introducing GeoVISTA Studio: An integrated suite of visualisation and computational methods for exploration and knowledge construction in geography. *Computers, Environment and Urban Systems, 26*(4), 267–292.

Geroimenko, V., & Chen, C. (2002). *Visualizing the semantic web XML-based internet and information visualisation.* Springer.

Leydesdorff, L. (1995). *The challenge of Scientometrics: The development, measurement, and self-organization of scientific communications.* Leiden: DSWO Press/Leiden University. Available from http://www.upublish.com/books/leydesdorff-sci.htm.

MacEachren, A. M., Wachowicz, M., Edsall, R., & Haug, D. (1999). Constructing knowledge from multivariate spatiotemporal data: Integrating geographic visualisation (GVis) with knowledge discovery in database (KDD) methods.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In L. M. Le Cam & J. Neyman (Eds.), *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability* (vol. 1, pp. 281–297). Berkeley, Califonia: University of California Press.

Mothe, J., Chrisment, C., Dousset, B., & Alaux, J. (2003). DocCube: Multi-dimensional visualisation and exploration of large document sets. *Proceedings of JASIST, 54*(7), 650–659.

Rijsbergen, K. Van (1979). *Information retrieval* (2nd ed.). London: Butterworths. Available from http://www.dcs.gla.ac.uk/Keith/Preface.html.

Skupin, A. (2004). The world of geography: Mapping a knowledge domain with cartographic means. *Proceedings of the National Academy of Sciences, 101*(Suppl. 1), 5274–5278. Available from http://www.pnas.org/cgi/content/full/101/suppl_1/5274.

Skupin, A., & Fabrikant, S. (2003). Spatialization methods: A cartographic research agenda for non-geographic information visualization. *Cartography and Geographic Information Science, 30*(2), 99–119.

Truong, Q. D. (2004). Visualisation de l'information. Research Master's Thesis.

Verbeek, A., Debackere, K., & Luwel, M. (2003). Science cited in patents: A geographic 'flow' analysis of bibliographic citation patterns in patents. *Scientometrics, 58*(2), 241–263.

Weigend, A. S., Weiner, E. D., & Peterson, J. O. (1999). Exploiting hierarchy on text categorization. *Information Retrieval Journal, I*(3), 193–216.

Widom, J. (1995). Research problems in data warehousing. In *International conference on information and knowledge management.*

White, H. D. (2003). Pathfinder networks and author co-citation analysis: A remapping of paradigmatic information scientists. *JASIST, 54*(5), 423–434.

White, H. D., & McCain, K. W. (1998). Visualizing a discipline: An author co-citation analysis of information science, 1972–1995. *JASIST, 49*(4), 327–355.

Zembowicz, R., & Zytkow, J. M. (1996). From contingency tables to various forms of knowledge in databases (Chapter 13). In U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, & R. Uthurusamy (Eds.), *Advances in knowledge discovery and data mining*, 0-262-56097-6. AAAI Press.

Zitt, M., & Bassecoulard, E. (1994). Development of a method for detection and trend analysis of research fronts built by lexical or co-citation analysis. *Scientometrics, 30*, 333–351.